ABSTRACT

In an ASP server farm, requests to use an application are directed to a particular executing instance of the application (or an appropriate component thereof) that is identified as being the least loaded of the available such instances of the application or its component. The number of such instances is dynamically increased or decreased in response to the number of requests for the application or components thereof. Requests may be directed (in accordance with the first aspect) or the instances adjusted (in accordance with a second aspect) on a per client-basis, in which instances of the application and/or components thereof are reserved for the use of a user or a particular group of users. Operation in this manner facilitates compliance with service agreements with respective users or groups of users. A record is maintained of the resources currently allocated and resources currently consumed for each combination of instance and user, in terms of load factors placed on the instances by requests serviced by those instances of the application or its components. The above arrangement permits a service agreement with a user to be structured in terms of an estimated range of requests for an application, with charges levied for servicing request rates within one or more different bands within this estimated range. Penalty rates are charged (in favour of the user) when requests are not serviced.